January 5, 2017 (Fictional Date)

From: Erica Junqueira

To: Whole Food Professionals

Re: Sales Trends and Forecasted Sales

Research Field: Sales Growth

Project Title: Whole Food Data

Executive Summary

This report contains an analysis of the net sales for Whole Foods from 1995 through 2016. The goal of this report is to identify any trends and seasonal patterns within the data to identify an Autoregressive Integrated Moving Average (ARIMA) model for the time-series data that can be used to forecast the net sales of the next 8 quarters.

Time Series analysis methods are conducted on the quarterly sales data. A natural log transformation is necessary to stabilize the increasing variance in the data. Various time series plots are used to identify trends and patterns in the data to determine the best model for forecasting net sales for the next two years.

The results of the time series analysis concluded that there is an upward trend in the data as well as seasonality. The net sales for the first quarter are higher on average than the other three quarters. An ARIMA (0,1,0,0,1,1,4) model is the best fit for the data. The model contains a first difference, fourth difference, and a moving average (MA) term. Differencing is needed to make the mean and variance constant over time. The first difference subtracts an observation from the observation that comes right before it, x_t-x_{t-1}. The seasonal difference is subtracting the quarter of one year to the corresponding quarter of the previous year, x_t-x_{t-4}. The MA term uses previous error values to predict current and future values.

The table below provides forecasted net sales for the next 8 quarters.

	Quarter1	**Quarter2**	**Quarter3**	**Quarter4**
2017	4,953,670	3,866,883	3,916,469	3,831,989
2018	5,428,198	4,237,304	4,291,641	4,199,068

Based on the forecasted sales, it is recommended to increase inventory as the sales continue to increase from one year to the next. It is important to increase inventory to account for the increased demand in quarter one. Additional staff hours are also recommended to keep shelves stocked and keep the lines at the checkout to a minimum. Additional analysis into regional data would be beneficial to provide more exact recommendations for individualized areas.

1.0 Project Description

Whole Foods is a supermarket chain in North America and the United Kingdom that focuses on selling organic healthy food options free from artificial colors, flavors, and preservatives. The client provided quarterly sales data from 1995-2016 that is analyzed to identify patterns to help predict future net sales for the company. The sales data will be used to look at the growth of the company over time and the possibility of seasonal sales trends. The overall trend and quarterly patterns in the net sales are important to understand so that sales projections can be created for the next two years (2017 and 2018).

1.1 Research Questions

On average, do the sales increase over the years for Whole Foods?

Food is a necessity all year long, but with holiday parties and summer barbeques, there may be a presence of quarterly sales trends. Is seasonality present in the Whole Foods data?

What are the sales projections for 2017 and 2018 predicted with 95% confidence?

1.2 Statistical Questions

Is there significant evidence to show an increase in sales data from one year to the next?

Is there a significant difference between the mean sales for the four quarters? Can we conclude that there is seasonality present in the data?

Can an ARIMA model be fit to the data that can be used to forecast sales for 2017 & 2018?

1.3 Variables of Interest

The analysis uses data provided by Whole Foods. The dataset contains sales data by quarter from 1995-2016. The variables of interest are Year, Net.Sales, and Q.

The floor function is used on Year to prepare it for analysis (i.e. 1995.75 became 1995). Year ranges from 1995-2016 with 4 observations per year. Net.Sales provides the total net sales for each quarter. Q is a categorical variable with 4 possible values, Q1, Q2, Q3, Q4. Lastly, there are four coded variables Q1-Q4 that use a (0,1) coding system.

2.0 Exploratory Data Analysis

The data from 1995-2016 contains 88 observations of quarterly sales data. Table 2.1.1 contains numerical summary data for Net.Sales. The net sales range in value from $161,864 to $4,829,000. A range this large suggests that the sales are continuing to increase year after year.

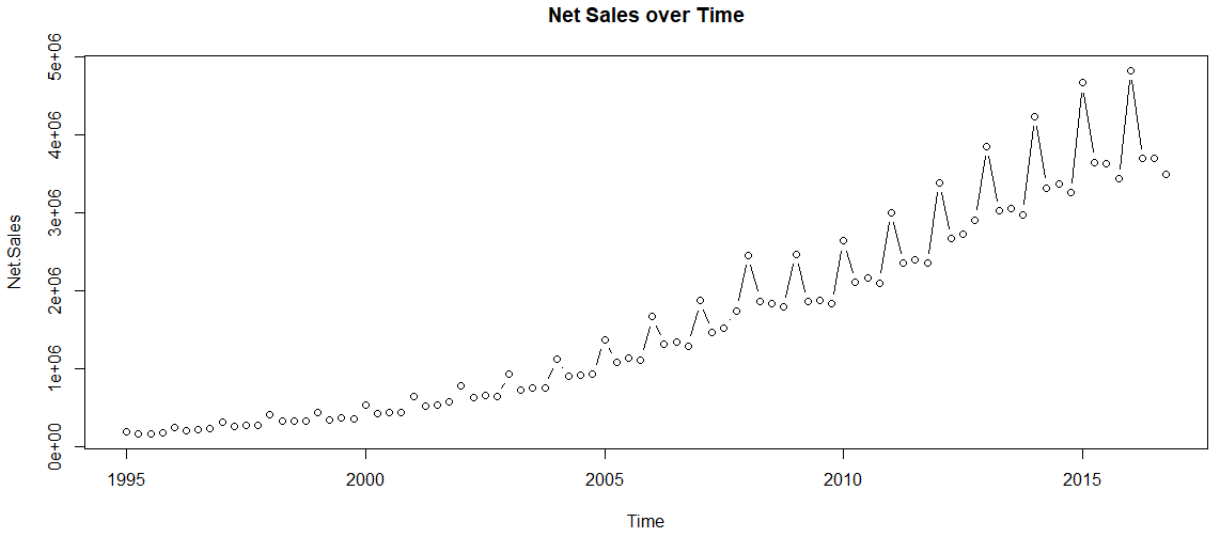
Table 2.1.1: Numerical Summary Data

Net.Sales
Min. : 161864
1st Qu.: 441501
Median :1301268
Mean :1605965
3rd Qu.:2509667
Max. :4829000

Figure 2.1.2 shows trend in the data with an increase in the net sales from year to year. There is a pattern that appears with the quarterly data. There appears to be a cluster of 3 quarters then a quarter that is higher than the others. A transformation of the data is needed to stabilize the variance in the data. There is an increase in variance as

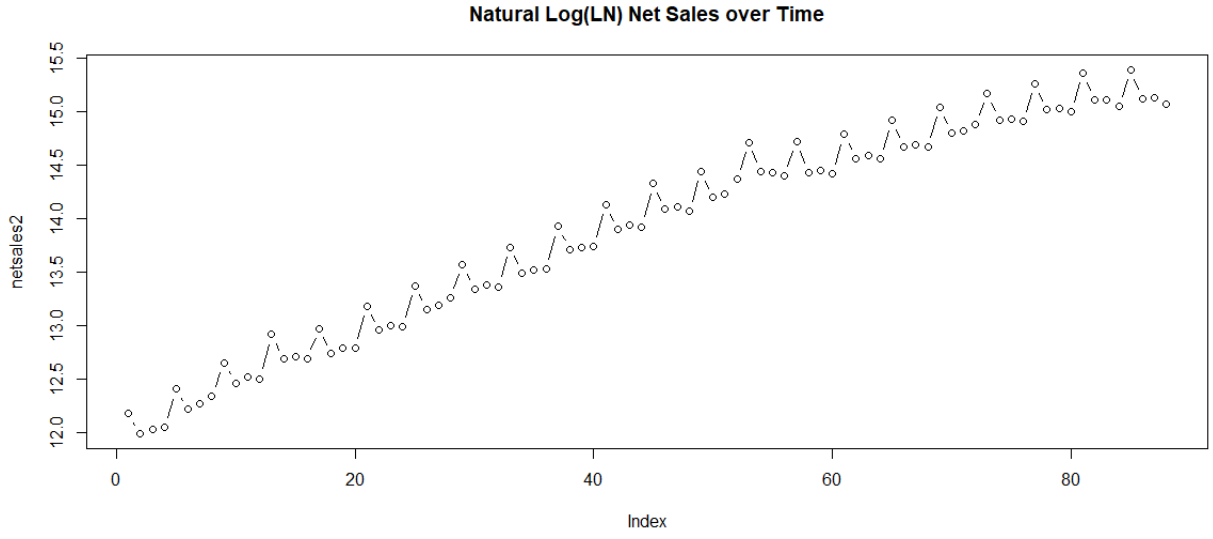
the data points spread out that is occurring as the mean increases. A natural log transformation is used to stabilize the variance. (i.e. ln(4,829,000)= 15.39015)

Figure 2.1.2: Plot of Net Sales over Time



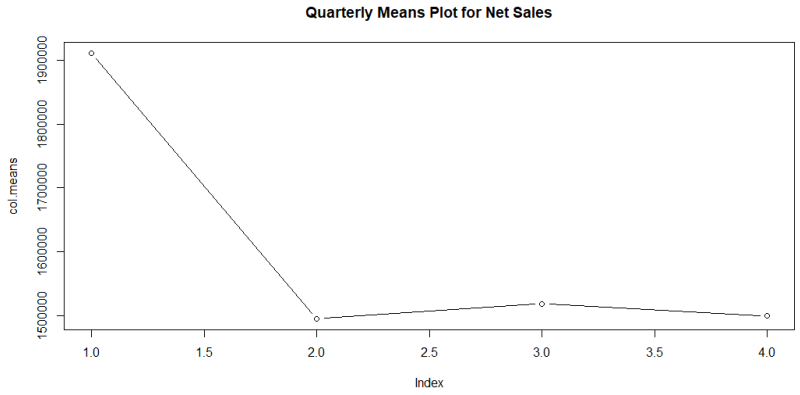
After the natural log of net sales is applied, the variance is stabilized. Figure 2.1.3 shows that there is no longer increasing variance over time. The seasonal pattern in the transformed data becomes even more clear. The transformed data will be used in the analysis.

Figure 2.1.3: Plot of the Natural Log Net Sales over Time



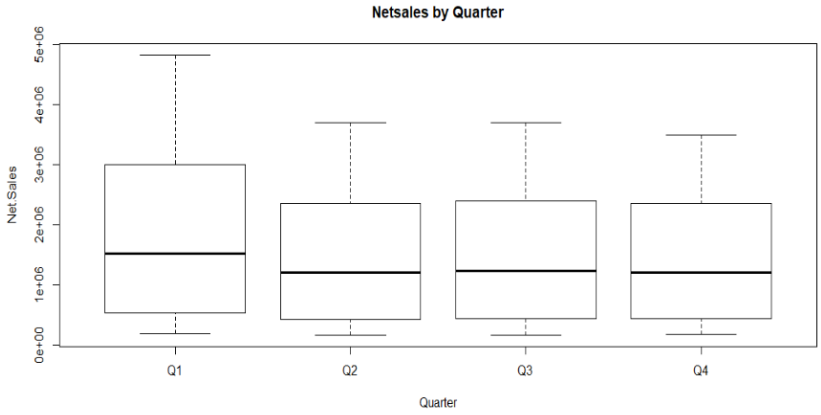
Further exploration of the quarterly data, Figure 2.1.4, shows that Quarter 1 has slightly higher net sales on average, approximately $1,900,000, than the other three quarters, approximately $1,500,000. The difference in the mean sales for quarter one does not appear to be significantly larger, but it may be large enough to show that seasonality is present in the data.

Figure 2.1.4: Boxplots of Net Sales by Quarter



A boxplot of the quarterly data, in Figure 2.1.5, shows that Quarter 1 has a median that is slightly higher than the other three quarters. The minimum net sales are about the same for all three, but Q1 has a maximum that is larger than the other three. Overall, the difference does not appear to be significant.

Figure 2.1.5: Boxplots of Net Sales by Quarter



In Table 2.1.6, numerical summaries are broken down by the sales by quarter. The mean net sales for Quarter 1 is $392,929 higher than the next largest average, Quarter 3.

Table 2.1.6: Numerical Summary Data by Quarter

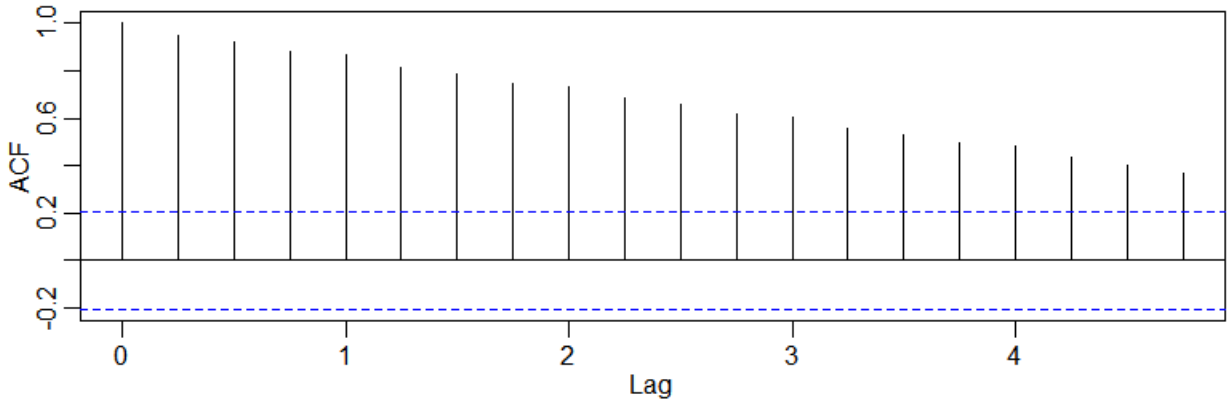
Q1	Q2	Q3	Q4
Net.Sales Min. : 195027 1st Qu.: 560328 Median :1517640 Mean :1911360 3rd Qu.:2912561 Max. :4829000	Net.Sales Min. : 161864 1st Qu.: 448000 Median :1198339 Mean :1494837 3rd Qu.:2289404 Max. :3696000	Net.Sales Min. : 167601 1st Qu.: 465814 Median :1235311 Mean :1518431 3rd Qu.:2340631 Max. :3703000	Net.Sales Min. : 171442 1st Qu.: 472888 Median :1203042 Mean :1499234 3rd Qu.:2289723 Max. :3497000

3.0 Statistical Analysis

3.1 Is there significant evidence to show an increase in sales data from one year to the next?

The time series plots in Figures 2.1.2 and 2.1.3 show that the net sales have an overall pattern of increasing year after year. Figure 3.1.1 of the autocorrelation function (ACF) of the transformed data confirms this is true. The lags of an ACF plot indicate a correlation in the data. For example, a spike at lag 4 would indicate that there is a correlation between values that are 4 time periods apart. Since most of the lags are close to 1, the ACF plot indicates that there is trend in the data. A first difference, $x_t - x_{t-1}$, is necessary to detrend the data. Since seasonality appears to be present based on the EDA above, it is important that the data does not become over differenced. To ensure that the data is not over differenced, the data was differenced for seasonality then the augmented Dickey-Fuller (ADF) test and the KPSS test were conducted to determine if trend was still present in the data. Both of the tests concluded that trend is present in the data, and a first difference is needed. (See Appendix 3.1.4)

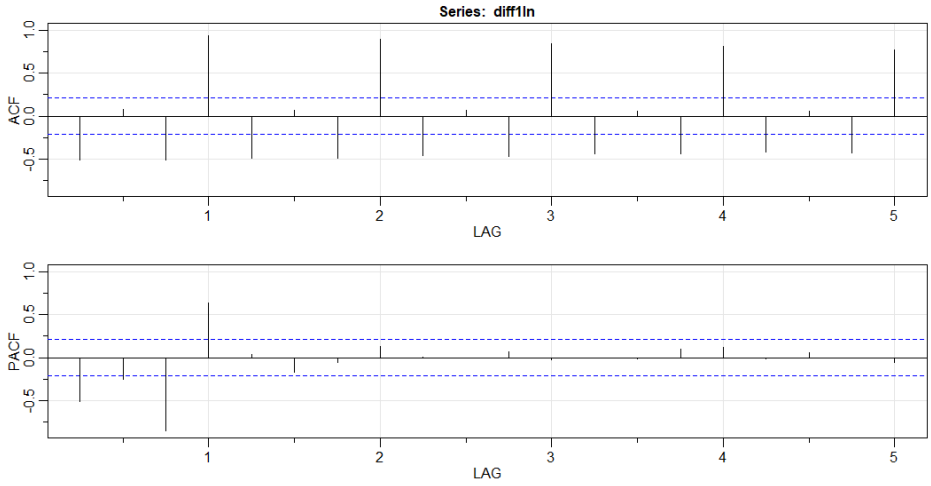
Figure 3.1.1: ACF for LN Net Sales



3.2 Is there a significant difference between the mean sales for the four quarters? Can we conclude that there is seasonality present in the data?

The EDA showed that the means sales for quarter 1 is larger than the other three quarters, indicating the presence of seasonality in the data. After concluding that a first difference is necessary to remove the trend in the data, the pattern of seasonality is easier to see in the ACF. The ACF of the detrended data, Figure 3.2.1, shows a clear pattern between the lags that are multiples of 4, indicating that the data is correlated to the time 4 periods prior. This confirms that there is a quarterly pattern in the data. The partial autocorrelation function (PACF) shows a couple of low-level spikes, but then just “shuts off”. To remove the seasonality in the data, the fourth difference is taken.

Figure 3.2.1: ACF and PACF after First Difference

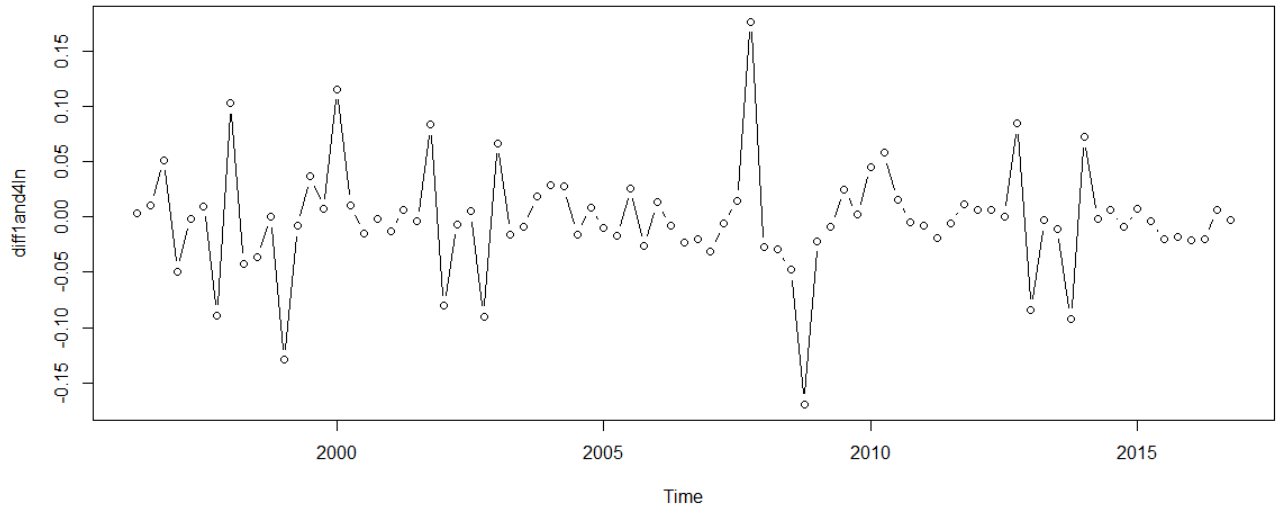


Regression analysis is conducted to identify significant differences in the mean sales by quarter. The tests can not conclude with 95% confidence that there is a difference between the mean sales for each quarter. (See Appendix 3.2.2) Seasonality is still present in the data, but average net sales by quarter are not significantly different. The numerical summaries in the EDA confirm this. There are differences in net sales, but the differences are not significant.

3.3 What are the sales projections for 2017 and 2018?

The plot in Figure 3.3.1 shows the net sales versus time after the first and fourth differences are taken on the data. Based on the plot, the trend and the seasonality has been removed from the data. A model will be fit to the differenced data.

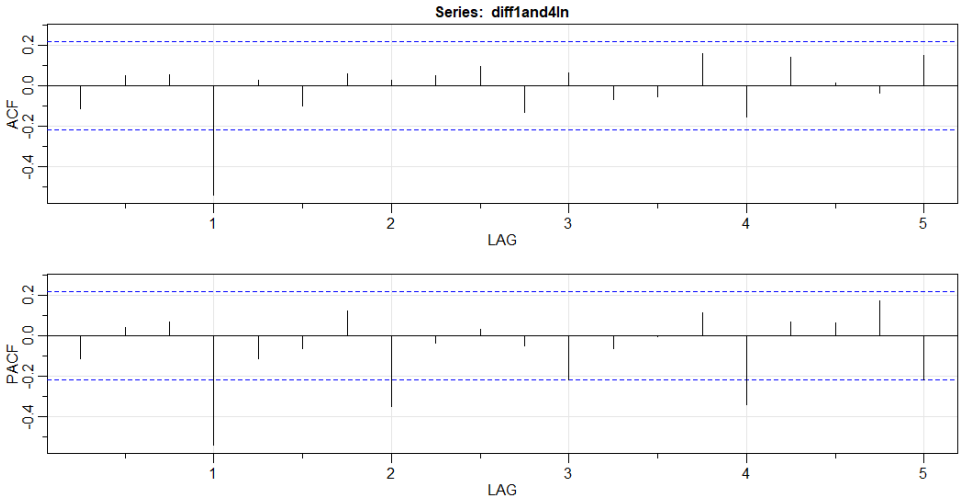
Figure 3.3.1: Times Series Plot after First and Fourth Differences



Now that the trend and seasonality are removed from the data, an ARIMA model can more easily be identified by looking at the ACF and PACF in Figure 3.3.2. The ARIMA model may contain autoregressive terms, differencing, and moving averages terms. It was identified earlier, that first and fourth differences are needed to account for the trend and seasonality in the data.

There are no significant spikes in the first three lags of the ACF or PACF, indicating that a non-seasonal term may not be needed. The ACF shows a significant lag 4, and the PACF shows a pattern in lags 4,8,12,16 indicating the possibility of a seasonal AR or MA term.

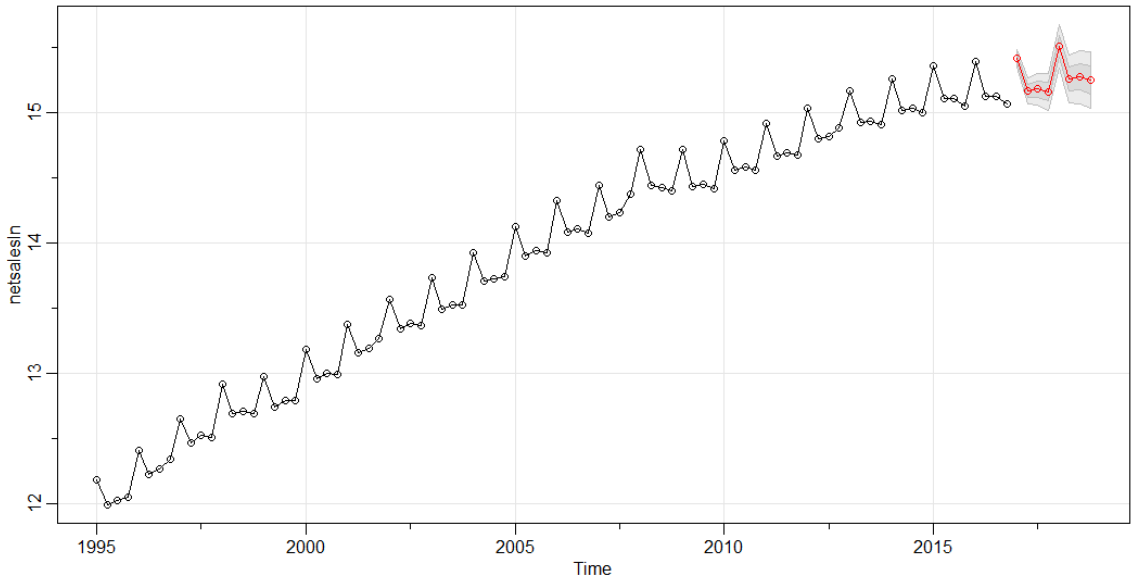
Figure 3.3.2: ACF and PACF after First and Fourth Differences



The best ARIMA model is (0,1,0,0,1,1,4). The first **1** in the model indicates the first difference, the second **1** indicates the seasonal difference, and the final **1** indicates a seasonal moving average (MA) term. MA terms use previous error values to predict current and future values. The summary results from the model can be found in Appendix 3.3.3. The seasonal MA term is significant at a 0.05 significance level. The residual plots in Appendix 3.3.4 were checked to make sure that the model assumptions were met. The summary results and residual plots indicate this model is a good fit for the data.

The model is then used to forecast the quarterly sales for 2017 and 2018. The values in red in Figure 3.3.5 predict the next 8 quarters. The forecasted values follow the same pattern of the original data indicating the model is an excellent fit. The y axis scale shows that the predictions are based on the natural log transformation of net sales.

Figure 3.3.5: Plot of Forecasted LN Net Sales for 2017 and 2018



The output for the net sales found in Appendix 3.3.6 is given in terms of the transformed data. Table 3.3.7 provides the predicted values in terms of dollar amounts to provide a more practical interpretation.

Table 3.3.7: Table of Predicted Values (to the nearest dollar)

	Quarter1	**Quarter2**	**Quarter3**	**Quarter4**
2017	4,953,670	3,866,883	3,916,469	3,831,989
2018	5,428,198	4,237,304	4,291,641	4,199,068

Table 3.3.8 contains 95% prediction intervals for the net sales predictions. The intervals provide a lower and upper bound for the net sales predictions. It can be concluded with 95% confidence that the net sales will fall between the lower and upper limits. The prediction intervals increase the farther out that the model is predicting because the standard error increases.

Table 3.3.8: 95% Confidence Intervals for the Predictions (to the nearest dollar)

		Quarter1	**Quarter2**	**Quarter3**	**Quarter4**
2017	lower limit	4,622,814	3,506,757	3,474,544	3,337,210
	upper limit	5,308,205	4,263,992	4,414,603	4,400,125
2018	lower limit	4,622,774	3,538,513	3,521,100	3,389,716
	upper limit	6,373,951	5,074,094	5,230,803	5,201,667

4.0 Recommendations

On average, do the sales increase from one year to the next for Whole Foods?

The results of the analysis concluded that there is an upward trend in the data. The net sales continue to increase year after year.

Is seasonality present in the Whole Foods data?

Seasonality is present in the data indicating that the net sales will depend on the quarter. From the analysis, the first quarter contains the highest net sales for each year, but the tests could not conclude with 95% confidence that the means are significantly different.

What are the sales projections for 2017 and 2018?

An ARIMA (0,1,0,0,1,1,4) model is fit to the net sales data to provide forecasted net sales for 2017 and 2018. The predicted sales followed the trend and seasonality found in the provided sales data. Table 4.1.1 below provides a quick comparison of the previous 2 years of data and the forecasted data for 2017 and 2018. It is recommended to increase inventory from the amount that was purchased in 2016. It is especially critical to account for the increasing demand in the first quarter. It is also recommended that Whole Foods increase staff hours to keep up with the stocking of shelves and longer lines.

Table 4.1.1: Comparison of Previous Two Years for Sales to Two Years of Forecasted Sales

	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2015	4,671,000	3,647,000	3,632,000	3,438,000
2016	4,829,000	3,696,000	3,703,000	3,497,000
2017	4,953,670	3,866,883	3,916,469	3,831,989
2018	5,428,198	4,237,304	4,291,641	4,199,068

5.0 Resources

R Studio was used to conduct the analysis, and the code will be provided upon request.

6.0 Considerations

This analysis only examines net sales by quarter from 1995-2016. Additional variables may be helpful for understanding the data. Providing data that includes the number of stores operating each year or the region stores are located in will help to provide a more thorough analysis and more accurate predictions.

The recommendations to increase inventory and staff hours, especially for the first quarter is based on this analysis that is limited to company-wide data. It does not analyze regions or individual stores. Further analysis of regional data is recommended, as it may indicate more specific needs to increase inventory and staff that may not be necessary across all of the stores. For example, regional data may also indicate different patterns of seasonality, perhaps based on areas with higher tourism or vacation houses.

It was an honor to work on this project for you. Please reach out with any additional questions that may arise.

Appendix

Appendix 3.1.2

The Augmented Dickey-Fuller Test and the KPSS tests are designed to test if a first difference is needed. The null hypothesis for the ADF test is the data are not stationary (requires a first difference). Testing with an a=0.05 significance level, since the p-value was greater (.08) than the significance level, the null hypothesis was concluded. A first difference is needed. The KPSS test the null hypothesis that the data is stationary and the alternative hypothesis that the data is not stationary (requires a first difference). Testing with an a=0.05 significance level, the p-value (0.01) is less than the significance level. The null hypothesis is rejected and the first difference is necessary.

Appendix 3.1.2 Results of Augmented Dickey- Fuller Test

```
        Augmented Dickey-Fuller Test

data:  diff4ln
Dickey-Fuller = -3.2807, Lag order = 4, p-value = 0.08035
alternative hypothesis: stationary
```

The p-value is greater than the 0.05 significance level. Therefore, we fail to reject the null hypothesis and conclude a first difference is needed.

Appendix 3.1.2 Results for KPSS Test

```
        KPSS Test for Level Stationarity

data:  diff4ln
KPSS Level = 1.1239, Truncation lag parameter = 3, p-value = 0.01
```

The p-value is less than the 0.05 significance level. Therefore, we reject the null hypothesis and conclude a first difference is needed.

Appendix 3.2.2

Linear Regression is conducted to determine if the means for the quarters are significantly different. Q4 is used as the reference level. The p-values for Q1, Q2, and Q3 are not significant. It is concluded that the means are not significantly different.

```
Call:
lm(formula = Net.Sales ~ Q1 + Q2 + Q3, data = quarterlyln)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8723 -0.8558  0.1602  0.8466  1.3370

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.8451758  0.2115173  65.456   <2e-16 ***
Q1           0.2079846  0.2991306   0.695    0.489
Q2          -0.0253030  0.2991306  -0.085    0.933
Q3           0.0003102  0.2991306   0.001    0.999
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9921 on 84 degrees of freedom
Multiple R-squared:  0.009365,  Adjusted R-squared:  -0.02602
F-statistic: 0.2647 on 3 and 84 DF,  p-value: 0.8507
```

The p-values are all larger than 0.05.

Appendix 3.3.3 Summary Results of Arima (0,1,0,0,1,1,4)

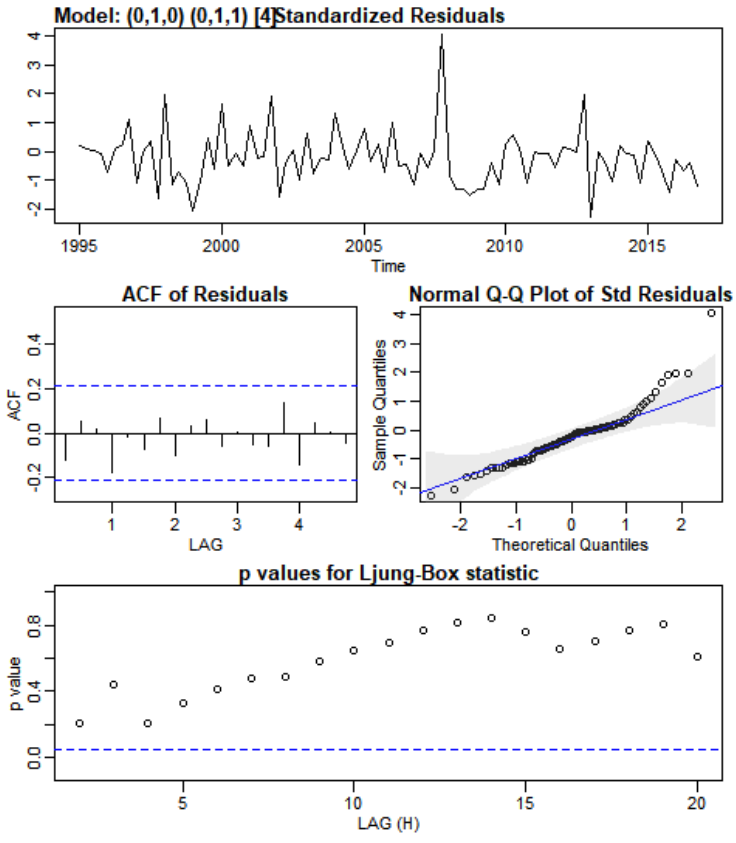
```
Coefficients:
          sma1
      -0.8174
s.e.   0.0698

sigma^2 estimated as 0.001244:  log likelihood = 157.64,  aic = -311.29

$degrees_of_freedom
[1] 82

$ttable
     Estimate     SE  t.value p.value
sma1  -0.8174 0.0698 -11.7036       0

$AIC
[1] -3.619593

$AICc
[1] -3.61904

$BIC
[1] -3.563341
```

The t-test indicates that the seasonal MA1 term is significant* with a corresponding p-value of zero.

*Significant at a=0.05

Appendix 3.3.4

Based on the residual plots all of the necessary assumptions of the model appear to be met. The standardized residuals do no show any clear patterns. The ACF of the residuals are not significant. The Q-Q plot does show one outlier, but there does not appear to be a significant violation of normality. The Ljung-Box show that none of the p-values are significant.



Appendix 3.3.6

```
$pred
           Qtr1        Qtr2        Qtr3        Qtr4
2017 15.41563921 15.16795928 15.18070109 15.15889462
2018 15.50711781 15.25943789 15.27217969 15.25037322

$se
              Qtr1          Qtr2          Qtr3          Qtr4
2017 0.03526787559 0.04987601990 0.06108528197 0.07053514007
2018 0.08194500008 0.09194930061 0.10096714677 0.10924310308
```

**The data for this analysis can be accessed at https://dasl.datadescription.com/datafile/whole-foods-2016/?_sfm_cases=4+59943&sf_paged=42.